

Singular Value Shrinkage Priors for Bayesian Prediction

Takeru MATSUDA and Fumiyasu KOMAKI

Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo
{Takeru_Matsuda, komaki}@mist.i.u-tokyo.ac.jp

Abstract

We develop singular value shrinkage priors for the mean matrix parameters in the matrix-variate Normal model with known covariance matrices. Introduced priors are superharmonic and put more weight on matrices with smaller singular values. They are a natural generalization of the Stein prior. Bayes estimators and Bayesian predictive densities based on introduced priors are minimax and dominate those based on the uniform prior in finite samples. The risk reduction is large when the true value of the parameter has small singular values. In particular, introduced priors perform stably well in the low rank case. We apply this result to multivariate linear regression problems by considering priors depending on the future samples. Introduced priors are compatible with reduced-rank regression.

1 Introduction

Suppose that we have a matrix observation $Y \sim N_{n,m}(M, C, \Sigma)$. Here we use the notation of Dawid (1981) for matrix-variate Normal distributions: $X \sim N_{n,m}(M, C, \Sigma)$ indicates that $X(n \times m)$ has a probability density

$$p(X) = \frac{1}{(2\pi)^{nm/2}(\det C)^{m/2}(\det \Sigma)^{n/2}} \text{etr} \left\{ -\frac{1}{2} \Sigma^{-1} (X - M)^\top C^{-1} (X - M) \right\}$$

with respect to the Lebesgue measure on $\mathbb{R}^{n \times m}$, where $\text{etr}(A) = \exp(\text{tr} A)$, $M(n \times m)$ is the mean matrix, $C(n \times n) \succ 0$ is the covariance matrix for rows, and $\Sigma(m \times m) \succ 0$ is the covariance matrix for columns. Here, we indicate the size of a matrix $A \in \mathbb{R}^{n \times m}$ by writing $A(n \times m)$. The vectorization $\text{vec}(X^\top)$ of X^\top satisfies

$$\text{vec}(X^\top) \sim N_{mn}(\text{vec}(M^\top), \Sigma \otimes C), \quad (1)$$

where \otimes denotes the Kronecker product, see Gupta & Nagar (2000). Here, the vectorization of $A(n \times m)$ is the $mn \times 1$ vector defined by

$$\text{vec}(A) = (a_{11}, \dots, a_{n1}, a_{12}, \dots, a_{n2}, \dots, a_{1m}, \dots, a_{nm})^\top,$$

and the Kronecker product $A \otimes B$ of two matrices $A(m \times n) = (a_{ij})$ and $B(p \times q) = (b_{ij})$ is the $mp \times nq$ matrix defined by

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

Matrix-variate Normal distributions appear naturally when we consider linear regression problems that have several dependent variables (multivariate linear regression).

We consider the prediction of $\tilde{Y} \sim N_{n,m}(M, \tilde{C}, \tilde{\Sigma})$ by a predictive density $\hat{p}(\tilde{Y}|Y)$. We evaluate predictive densities by the Kullback–Leibler divergence

$$D(\tilde{p}(\cdot|M), \hat{p}(\cdot|Y)) = \int \tilde{p}(\tilde{Y}|M) \log \frac{\tilde{p}(\tilde{Y}|M)}{\hat{p}(\tilde{Y}|Y)} d\tilde{Y}$$

as a loss function. The Kullback–Leibler risk function of a predictive density $\hat{p}(\tilde{Y}|Y)$ is

$$\begin{aligned} R_{\text{KL}}(M, \hat{p}) &= E[D\{\tilde{p}(\cdot|M), \hat{p}(\cdot|Y)\}] \\ &= \int \int p(Y|M) \tilde{p}(\tilde{Y}|M) \log \frac{\tilde{p}(\tilde{Y}|M)}{\hat{p}(\tilde{Y}|Y)} dY d\tilde{Y}. \end{aligned}$$

We consider Bayesian predictive densities with a prior $\pi(M)$:

$$\begin{aligned} \hat{p}_\pi(\tilde{Y}|Y) &= \int \tilde{p}(\tilde{Y}|M) \pi(M|Y) dM \\ &= \frac{\int \tilde{p}(\tilde{Y}|M) p(Y|M) \pi(M) dM}{\int \tilde{p}(\tilde{Y}|M) \pi(M) dM}. \end{aligned}$$

In the following, we assume that $n - m \geq 2$. When $m = 1$, this problem reduces to the prediction of vector-variate Normal model $y \sim N_n(\mu, \Sigma)$, $\tilde{y} \sim N_n(\mu, \tilde{\Sigma})$. This problem has been studied by several authors under the assumption that $n \geq 3$. First, the cases where $\tilde{\Sigma}$ is proportional to Σ were investigated. Komaki (2001) gave the analytical form of Bayesian predictive densities based on the Stein prior $\pi_S(\mu) = \|\mu\|^{-(n-2)}$ and proved that Bayesian predictive densities based on the Stein prior dominate those based on the Jeffreys prior under the Kullback–Leibler risk. Since the Stein prior puts more weight near the origin, the amount of risk reduction is large when

the true value of μ is near the origin. George et al. (2006) generalized this result and proved that Bayesian predictive densities based on superharmonic priors dominate those based on the Jeffreys prior under the Kullback–Leibler risk. Next, Kobayashi & Komaki (2008) and George & Xu (2008) considered the cases where $\tilde{\Sigma}$ is not necessarily proportional to Σ . Bayesian predictive densities based on superharmonic priors dominate those based on the uniform prior under the Kullback–Leibler risk also in this general situation. Kobayashi & Komaki (2008) and George & Xu (2008) applied their results to linear regression problems.

Now, since matrix-variate Normal distributions are special cases of vector-variate Normal distributions as in (1), the above results also apply to the matrix-variate Normal distributions. In this paper, we propose superharmonic priors that shrink the singular values of the mean matrix parameters M . From the results of Stein (1974), Bayes estimators based on the introduced priors dominate the maximum likelihood estimator and are thus minimax. Furthermore, Bayesian predictive densities based on the introduced priors dominate those based on the uniform prior and are thus minimax. The amount of risk reduction is large when the true value of the mean matrix parameter M has smaller singular values. Therefore, introduced priors work particularly well when the true value of the mean matrix parameter M is low rank, since low rank matrices have sparse singular values. We apply the introduced priors to the multivariate linear regression problems, where we can reasonably expect the low rank assumption to be satisfied from the theory of reduced rank regression (Reinsel & Velu, 1998). Finally, we provide some numerical results.

This paper is organized as follows. In Section 2, we propose singular value shrinkage priors, prove that they are superharmonic and provide the analytical form of Bayesian predictive densities based on them. In Section 3, we apply singular value shrinkage priors to multivariate linear regression problems and discuss the relationship with reduced-rank regression. In Section 4, some numerical results are provided. Concluding remarks are given in Section 5.

2 Singular value shrinkage priors

2.1 Singular value shrinkage estimators

The idea of singular value shrinkage was utilized by Efron & Morris (1972) and Stein (1974). Here we review their work. In this subsection, we assume $X \sim N_{n,m}(M, I, I)$.

Let

$$X = U\Sigma V^\top,$$

$$U \in O(n), \quad V \in O(m), \quad \Sigma = (\text{diag}(\sigma_1, \dots, \sigma_m) \quad O_{m,n-m})^\top,$$

and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$$

be the singular value decomposition of a matrix X , where $O_{m,n-m}$ is the zero matrix of size $m \times (n-m)$, and $\sigma_1, \dots, \sigma_m$ are the singular values of X . Similarly, let

$$\hat{M} = \hat{U} \hat{\Sigma} \hat{V}^\top,$$

$$\hat{U} \in O(n), \quad \hat{V} \in O(m), \quad \hat{\Sigma} = (\text{diag}(\hat{\sigma}_1, \dots, \hat{\sigma}_m) \quad O_{m,n-m})^\top,$$

and

$$\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_m \geq 0$$

be the singular value decomposition of an estimator \hat{M} of M , where $\hat{\sigma}_1, \dots, \hat{\sigma}_m$ are the singular values of \hat{M} .

Efron & Morris (1972) proposed

$$\hat{M}_{\text{EM}} = X \{I_m - (n - m - 1)S^{-1}\}$$

as an empirical Bayes estimator, where $S = X^\top X$. They proved that \hat{M}_{EM} is minimax and dominates the maximum likelihood estimator under the Frobenius loss

$$l(M, \hat{M}) = \|\hat{M} - M\|_{\text{F}}^2 = \sum_{i,j} (\hat{M}_{ij} - M_{ij})^2.$$

Stein (1974) noticed that \hat{M}_{EM} can be represented in singular value decomposition form as

$$\hat{\sigma}_i = \left(1 - \frac{n - m - 1}{\sigma_i^2}\right) \sigma_i \quad (i = 1, \dots, m),$$

$$\hat{U} = U, \quad \hat{V} = V.$$

Therefore, \hat{M}_{EM} shrinks the singular values of the observation X . When $m = 1$, \hat{M}_{EM} coincides with the James-Stein estimator.

We note that \hat{M}_{EM} is not a Bayes estimator. In this study, we develop priors shrinking the singular values. The Bayes estimator based on the introduced prior has properties similar to those of \hat{M}_{EM} . This is an extension of the relationship between the James-Stein estimator and the Stein prior.

2.2 Definition of the singular value shrinkage prior and its superharmonicity

We consider a prior defined by

$$\pi_{\text{SVS}}(M) = \det(M^\top M)^{-(n-m-1)/2}. \quad (2)$$

Note that we assume $n - m \geq 2$. From the relation

$$\det(M^\top M) = \prod_{i=1}^m \sigma_i(M)^2,$$

where $\sigma_i(M)$ denotes the i th singular value of M , we obtain

$$\pi_{\text{SVS}}(M) = \prod_{i=1}^m \sigma_i(M)^{-(n-m-1)}.$$

Therefore, this prior puts more weight on matrices with smaller singular values. When $m = 1$, π_{SVS} coincides with the Stein prior $\pi_{\text{S}}(\mu) = \|\mu\|^{-(n-2)}$. We call π_{SVS} the singular value shrinkage prior in the following.

We provide a proof of superharmonicity of the singular value shrinkage prior π_{SVS} . An extended real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$ is said to be superharmonic if it satisfies the following properties (Helms, 2009, p. 70):

1. f is lower semicontinuous.
2. $f \not\equiv \infty$.
3. $\frac{1}{\Omega_d \delta^{d-1}} \int_{S_{x,\delta}} f(z) ds(z) \leq f(x)$ for every $x \in \mathbb{R}^d$ and $\delta > 0$, where Ω_d denotes the surface area of the unit sphere in \mathbb{R}^d and $S_{x,\delta}$ denotes the sphere with center x and radius r .

If f is a C^2 function, then f is superharmonic if and only if

$$\Delta f(x) = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2} f(x) \leq 0$$

holds for every x from Lemma 3.3.4 of Helms (2009).

We define a function $f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ to be superharmonic when $f \circ \text{vec} : \mathbb{R}^{mn} \rightarrow \mathbb{R}$ is superharmonic.

Theorem 1. *The prior density π_{SVS} is superharmonic.*

Proof. First we prove that $\det(M^\top M + \varepsilon I_m)^{-(n-m-1)/2}$ is superharmonic for every $\varepsilon > 0$.

We write the (i, j) th entry of a matrix X by X_{ij} and the (i, j) th entry of X^{-1} by X^{ij} . Let

$$K = M^\top M + \varepsilon I,$$

so that

$$K_{ab} = \sum_i M_{ia} M_{ib} + \varepsilon \delta_{ab} = K_{ba},$$

where the subscripts a, b, \dots run from 1 to m and the subscripts i, j, \dots run from 1 to n , δ_{ab} is 1 when $a = b$ and 0 otherwise, and the (i, j) th entry of a matrix X is denoted by X_{ij} . From the definition,

$$\frac{\partial K_{bc}}{\partial M_{ia}} = \delta_{ac} M_{ib} + \delta_{ab} M_{ic}. \quad (3)$$

By using

$$\frac{\partial}{\partial K_{ab}} \det K = K^{ab} \det K$$

and (3), we obtain

$$\frac{\partial}{\partial M_{ia}} \det K = \sum_{b,c} \frac{\partial K_{bc}}{\partial M_{ia}} \frac{\partial}{\partial K_{bc}} \det K = 2 \sum_b M_{ib} K^{ab} \det K, \quad (4)$$

where K^{ab} is the (a, b) th entry of the inverse matrix of K^{-1} . Therefore,

$$\begin{aligned} \frac{\partial^2}{\partial M_{ia}^2} \det K &= 2 \left(\sum_b \delta_{ab} K^{ab} \right) \det K + 2 \left(\sum_b M_{ib} \frac{\partial K^{ab}}{\partial M_{ia}} \right) \det K \\ &\quad + 2 \left(\sum_b M_{ib} K^{ab} \right) \frac{\partial}{\partial M_{ia}} \det K \\ &= 2 K^{aa} \det K - 2 \left(\sum_{b,d} M_{ib} M_{id} K^{aa} K^{bd} \right) \det K \\ &\quad - 2 \left(\sum_{b,c} M_{ib} M_{ic} K^{ab} K^{ac} \right) \det K + 4 \left(\sum_{b,c} M_{ib} M_{ic} K^{ab} K^{ac} \right) \det K. \end{aligned} \quad (5)$$

Here, we used

$$\frac{\partial K^{ab}}{\partial M_{ia}} = - \sum_c M_{ic} K^{aa} K^{bc} - \sum_d M_{id} K^{ad} K^{ab},$$

which can be derived by differentiating the equation $\sum_c K_{bc} K^{ac} = \delta_{ab}$.

We have

$$\begin{aligned} \Delta(\det K)^{-(n-m-1)/2} &= \sum_{i,a} \frac{\partial^2}{\partial M_{ia}^2} (\det K)^{-(n-m-1)/2} \\ &= \frac{n-m-1}{2} (\det K)^{-(n-m-1)/2} \sum_{i,a} (A_{ia} + B_{ia}), \end{aligned} \quad (6)$$

where

$$A_{ia} = \frac{n-m+1}{2} (\det K)^{-2} \left(\frac{\partial}{\partial M_{ia}} \det K \right)^2,$$

and

$$B_{ia} = -(\det K)^{-1} \frac{\partial^2}{\partial M_{ia}^2} \det K.$$

By using (4), we obtain

$$A_{ia} = 2(n-m+1) \left(\sum_b M_{ib} K^{ab} \right) \left(\sum_c M_{ic} K^{ac} \right).$$

Thus,

$$\begin{aligned} \sum_i A_{ia} &= 2(n-m+1) \sum_{b,c} \left(K^{ab} K^{ac} \sum_i M_{ib} M_{ic} \right) \\ &= 2(n-m+1) \sum_{b,c} K^{ab} K^{ac} (K_{bc} - \varepsilon \delta_{bc}) \\ &= 2(n-m+1) \left(K^{aa} - \varepsilon \sum_b K^{ab} K^{ab} \right). \end{aligned}$$

On the other hand, by (5),

$$\sum_i B_{ia} = -2(n-m+1) K^{aa} - 2\varepsilon K^{aa} \sum_b K^{bb} + 2\varepsilon \sum_b K^{ab} K^{ab}.$$

Hence, noting that $K^{bb} = \sum_i M_{ib}^2 + \varepsilon > 0$, we obtain

$$\sum_i (A_{ia} + B_{ia}) = -2(n-m)\varepsilon \sum_b K^{ab} K^{ab} - 2\varepsilon K^{aa} \sum_b K^{bb} < 0. \quad (7)$$

Thus, from (6) and (7),

$$\Delta(\det K)^{-(n-m-1)/2} < 0. \quad (8)$$

Therefore, $\det(M^\top M + \varepsilon I)^{-(n-m-1)/2}$ is superharmonic for every $\varepsilon > 0$.

Now, let

$$\pi^{(k)}(M) = \det \left(M^\top M + \frac{1}{k} I \right)^{-(n-m-1)/2}.$$

Then, $\pi^{(k)}$ is superharmonic by (8) and $\pi^{(1)} \leq \pi^{(2)} \leq \dots$, since

$$\pi^{(k)}(M) = \prod_{i=1}^m \left\{ \lambda_i(M^\top M) + \frac{1}{k} \right\}^{-(n-m-1)/2},$$

where $\lambda_i(M^\top M) \geq 0$ denotes the i th eigenvalue of $M^\top M$. Also,

$$\lim_{k \rightarrow \infty} \pi^{(k)}(M) = \pi_{\text{SVS}}(M)$$

for every M . Therefore, by Theorem 3.4.8 of Helms (2009), π_{SVS} is superharmonic. \square

From this proof, we immediately obtain the following Theorem.

Theorem 2. *The prior π_{SVS} satisfies $\Delta\pi_{\text{SVS}}(M) = 0$ if the rank of M is full.*

Proof. Consider $\varepsilon \rightarrow 0$ in (7). \square

Therefore, the superharmonicity of π_{SVS} is strongly concentrated in the same way as the Laplacian of the Stein prior becomes a Dirac delta function.

Another interesting point about this prior is that π_{SVS} is superharmonic column-wise: π_{SVS} is superharmonic as a function of the i th column of M for every M ($i = 1, \dots, m$).

We provide another proof of Theorem 2 in the Appendix.

2.3 Minimavity of Bayes estimators and Bayesian predictive densities based on π_{SVS}

We prove that Bayesian predictive densities and Bayes estimators based on the singular value shrinkage priors are minimax.

When $X \sim N_{n,m}(M, C, \Sigma)$, we define the marginal distribution of X with prior $\pi(M)$ by

$$m_\pi(X; C, \Sigma) = \int p(X|M, C, \Sigma)\pi(M)dM.$$

When $\pi = \pi_{\text{SVS}}$, we denote the marginal distribution by m_{SVS} .

Lemma 1. *If $m_\pi(X; C, \Sigma) < \infty$ for every X and π is superharmonic, then $m_\pi(X; C, \Sigma)$ is superharmonic.*

Proof. Let

$$\phi(X; C, \Sigma) = p(X|O, C, \Sigma).$$

Then, putting $A = X - M$,

$$m_\pi(X; C, \Sigma) = \int \phi(X - M; C, \Sigma)\pi(M)dM = \int \phi(A; C, \Sigma)\pi(X - A)dA.$$

Now, for every $x \in \mathbb{R}^d$ and $\delta > 0$,

$$\begin{aligned} & \frac{1}{\Omega_d \delta^{d-1}} \int_{S_{x,\delta}} m_\pi(X; C, \Sigma) ds(X) \\ &= \frac{1}{\Omega_d \delta^{d-1}} \int_{S_{x,\delta}} \left\{ \int \phi(A; C, \Sigma) \pi(X - A) dA \right\} ds(X) \\ &= \frac{1}{\Omega_d \delta^{d-1}} \int \phi(A; C, \Sigma) \left\{ \int_{S_{x,\delta}} \pi(X - A) ds(X) \right\} dA \\ &\leq \frac{1}{\Omega_d \delta^{d-1}} \int \phi(A; C, \Sigma) \pi(-A) dA = m_\pi(O). \end{aligned}$$

Here, the second equation follows from Fubini's Theorem and the third inequality follows from the superharmonicity of π . Therefore, $m_\pi(X; C, \Sigma)$ is superharmonic. \square

In terms of estimation of M from $X \sim N(M, C, \Sigma)$, the following result (Stein, 1974) is known.

Lemma 2. (Stein, 1974) *If m_π satisfies $\Delta m_\pi(X; C, \Sigma) \leq 0$, then Bayes estimator \hat{M}^π with prior π is minimax under the Frobenius loss. Furthermore, Bayes estimator with prior π dominates the maximum likelihood estimator unless π is the uniform prior.*

Lemma 2 (Stein (1974)) is obtained from the expression

$$\begin{aligned} & \mathbb{E}_M \left[\|\hat{M}^{\text{MLE}} - M\|_F^2 \right] - \mathbb{E}_M \left[\|\hat{M}^\pi - M\|_F^2 \right] \\ &= \mathbb{E}_M \left[\|\nabla \log m_\pi(X)\|^2 - 2 \frac{\Delta m_\pi(X)}{m_\pi(X)} \right]. \end{aligned} \quad (9)$$

of the risk difference between the maximum likelihood estimator and Bayes estimator with prior π .

In terms of prediction, we obtain the following result. Here, we consider the case where $\tilde{\Sigma} \otimes \tilde{C}$ is proportional to $\Sigma \otimes C$. This assumption corresponds to the setting of Komaki (2001) and George et al. (2006). The general covariance case is considered in Section 2.5. We assume $C = v_1 I_n, \tilde{C} = v_2 I_n, \Sigma = \tilde{\Sigma} = I_m$ without loss of generality. Let

$$v_0 = \frac{v_1 v_2}{v_1 + v_2} (< v_1).$$

We write Bayesian predictive density based on the uniform prior $\pi_I(M) = 1$ by $\hat{p}_I(\tilde{Y}|Y)$.

Proposition 1. *If π is superharmonic, then $\hat{p}_\pi(\tilde{Y}|Y)$ is minimax under the Kullback–Leibler risk R_{KL} . Furthermore, $\hat{p}_\pi(\tilde{Y}|Y)$ dominates $\hat{p}_I(\tilde{Y}|Y)$ unless π is the uniform prior.*

Proof. From Lemma 2 of George et al. (2006), the difference of R_{KL} is

$$\begin{aligned} & R_{\text{KL}}(M, \hat{p}_I) - R_{\text{KL}}(M, \hat{p}_\pi) \\ &= \mathbb{E}_{M, v_0} \log m_\pi(Z; v_0 I_n, I_m) - \mathbb{E}_{M, v_1} \log m_\pi(Z; v_1 I_n, I_m), \end{aligned}$$

where $\mathbb{E}_{M, v}(\cdot)$ denotes the expectation with respect to $Z \sim N_{n, m}(M, v I_n, I_m)$. It can be rewritten as

$$\begin{aligned} & R_{\text{KL}}(M, \hat{p}_I) - R_{\text{KL}}(M, \hat{p}_\pi) \\ &= \mathbb{E}_{M, v_0} \{ \log m_\pi(Z; v_0 I_n, I_m) - \log m_\pi(Z; v_1 I_n, I_m) \} \\ & \quad + \{ \mathbb{E}_{M, v_0} \log m_\pi(Z; v_1 I_n, I_m) - \mathbb{E}_{M, v_1} \log m_\pi(Z; v_1 I_n, I_m) \}. \end{aligned} \quad (10)$$

Since π is superharmonic, from Lemma 3.4.4 of Helms (2009) and $v_0 < v_1$,

$$\begin{aligned} m_\pi(Z; v_0 I_n, I_m) &= \mathbb{E}_{M \sim N_{n,m}(Z, v_0 I_n, I_m)} \pi(M) \\ &\geq \mathbb{E}_{M \sim N_{n,m}(Z, v_1 I_n, I_m)} \pi(M) = m_\pi(Z; v_1 I_n, I_m). \end{aligned}$$

Thus, the first term of the right hand side of (10) is nonnegative. Also, since m_π is superharmonic from Lemma 1 and logarithm of a superharmonic function is superharmonic, $\log m_\pi(Z; v_1 I_n, I_m)$ is superharmonic. Then, from Lemma 3.4.4 of Helms (2009) and $v_0 < v_1$, the second term of the right hand side of (10) is nonnegative. Therefore, (10) is nonnegative. \square

Proposition 1 is without an assumption concerning twice differentiability of π and is a slight generalization of the results in George et al. (2006).

From Theorem 1 and Lemma 1, m_{SVS} is superharmonic. Also, from (14) in the next subsection, m_{SVS} is a C^2 function. Therefore, the singular value shrinkage prior π_{SVS} satisfies the conditions of Lemma 2 and Proposition 1.

By combining the results above, we obtain the following.

Theorem 3.

1. *The Bayes estimator based on the prior π_{SVS} is minimax and dominates the maximum likelihood estimator under the Frobenius risk.*
2. *The Bayesian predictive density $\hat{p}_{\text{SVS}}(\tilde{Y}|Y)$ based on the prior π_{SVS} is minimax and dominates $\hat{p}_1(\tilde{Y}|Y)$ under the Kullback–Leibler risk R_{KL} .*

Since π_{SVS} shrinks the singular values for each, the risk reduction of π_{SVS} is large when the true value of the parameter M has small singular values. A remarkable point is that π_{SVS} works well even when only part of the singular values are small. In particular, π_{SVS} works stably well in the low rank case, which indicates the sparsity of singular values. We confirm these facts by numerical experiments in the next section.

2.4 The Bayesian predictive density based on π_{SVS}

We provide the analytical form of Bayesian predictive densities based on the singular value shrinkage priors. Here, we consider the case where $\tilde{\Sigma} \otimes \tilde{C}$ is proportional to $\Sigma \otimes C$, and assume $C = \tilde{C} = I_n, \Sigma = \tilde{\Sigma} = I_m$ without loss of generality.

Theorem 4. *The Bayesian predictive density based on the singular value shrinkage prior (2) is*

$$\begin{aligned} \hat{p}_{\text{SVS}}(\tilde{Y}|Y) &= (2\pi)^{-\frac{1}{2}mn} \text{etr} \left\{ -\frac{1}{4}(\tilde{Y} - Y)^\top (\tilde{Y} - Y) - \frac{1}{4}Z^\top Z + \frac{1}{2}Y^\top Y \right\} \\ &\quad \times \frac{2^{-\frac{1}{2}m(m+1)} {}_1F_1\left(\frac{m+1}{2}; \frac{n}{2}; \frac{1}{4}Z^\top Z\right)}{{}_1F_1\left(\frac{m+1}{2}; \frac{n}{2}; \frac{1}{2}Y^\top Y\right)}. \end{aligned} \quad (11)$$

Here, $Z = Y + \tilde{Y}$ and ${}_pF_q$ is the hypergeometric function of matrix argument (Gupta & Nagar, 2000, p. 34) defined by

$${}_pF_q(a_1, \dots, a_p; b_1, \dots, b_q; S) = \sum_{k=0}^{\infty} \sum_{\kappa \vdash k} \frac{(a_1)_{\kappa} \cdots (a_p)_{\kappa}}{(b_1)_{\kappa} \cdots (b_q)_{\kappa}} \frac{C_{\kappa}(S)}{k!},$$

where a_i, b_j are arbitrary complex numbers, S is a $p \times p$ complex symmetric matrix, $\sum_{\kappa \vdash k}$ denotes summation over all partitions κ of k , $(a)_{\kappa}$ is the generalized Pochhammer symbol, and $C_{\kappa}(S)$ denotes the Zonal polynomial.

Proof. We represent the Bayesian predictive density as

$$\hat{p}_{\pi}(\tilde{Y}|Y) = \frac{m_{\pi}(Y, \tilde{Y})}{m_{\pi}(Y)} = \frac{m_{\pi}(Y, \tilde{Y})}{m_{\pi}(Z)} \frac{m_{\pi}(Z)}{m_{\pi}(Y)}, \quad (12)$$

where m_{π} denotes the marginal distribution with prior π . Here, $m_{\pi}(Y, \tilde{Y})/m_{\pi}(Z)$ does not depend on π , because Z is the sufficient statistic for M . Therefore, we obtain

$$\begin{aligned} \frac{m_{\pi}(Y, \tilde{Y})}{m_{\pi}(Z)} &= \frac{p(Y, \tilde{Y}|M)}{p(Z|M)} \\ &= (2\pi)^{-mn} \text{etr} \left\{ -\frac{1}{2}(Y - M)^{\top}(Y - M) - \frac{1}{2}(\tilde{Y} - M)^{\top}(\tilde{Y} - M) \right\} \\ &\quad \times (4\pi)^{\frac{1}{2}mn} \text{etr} \left\{ \frac{1}{4}(Z - 2M)^{\top}(Z - 2M) \right\} \\ &= \pi^{-\frac{1}{2}mn} \text{etr} \left\{ -\frac{1}{4}(\tilde{Y} - Y)^{\top}(\tilde{Y} - Y) \right\}. \end{aligned} \quad (13)$$

Next, we calculate $m_{\text{SVS}}(Y)$ and $m_{\text{SVS}}(Z)$. From the definition, $m_{\text{SVS}}(Y)$ is

$$m_{\text{SVS}}(Y) = \int p(Y|M) \pi(M) dM.$$

Now we interpret $p(Y|M)$ as the probability density of $M \sim N_{n,m}(Y, I, I)$. Then $m_{\text{SVS}}(Y)$ is viewed as the expectation of $\pi(M)$. From Theorem 3.5.1 in Gupta & Nagar (2000), $S = M^{\top}M$ has a noncentral Wishart distribution:

$$S \sim W_m(n, I_m, Y^{\top}Y).$$

Therefore,

$$m_{\text{SVS}}(Y) = \text{E} \left[(\det S)^{-(n-m-1)/2} \right].$$

By using Theorem 3.5.6 in Gupta & Nagar (2000), we obtain

$$\begin{aligned} &\text{E} \left[(\det S)^{-(n-m-1)/2} \right] \\ &= \frac{2^{-\frac{m(n-m-1)}{2}} \Gamma_m\left(\frac{m+1}{2}\right)}{\Gamma_m\left(\frac{n}{2}\right)} \text{etr} \left(-\frac{1}{2}Y^{\top}Y \right) {}_1F_1 \left(\frac{m+1}{2}; \frac{n}{2}; \frac{1}{2}Y^{\top}Y \right). \end{aligned}$$

Using this, we obtain

$$m_{\text{SVS}}(Y) = \frac{2^{-\frac{m(n-m-1)}{2}} \Gamma_m\left(\frac{m+1}{2}\right)}{\Gamma_m\left(\frac{n}{2}\right)} \text{etr}\left(-\frac{1}{2}Y^\top Y\right) \times {}_1F_1\left(\frac{m+1}{2}; \frac{n}{2}; \frac{1}{2}Y^\top Y\right). \quad (14)$$

Similarly, we obtain

$$m_{\text{SVS}}(Z) = \frac{2^{-mn} \Gamma_m\left(\frac{m+1}{2}\right)}{\Gamma_m\left(\frac{n}{2}\right)} \text{etr}\left(-\frac{1}{4}Z^\top Z\right) {}_1F_1\left(\frac{m+1}{2}; \frac{n}{2}; \frac{1}{4}Z^\top Z\right). \quad (15)$$

Substituting (13), (14), and (15) to (12), we obtain the Theorem. \square

2.5 Singular value shrinkage priors depending on the future covariance matrices

Thus far, we assumed that the covariance matrices of observation data C and Σ and of future data \tilde{C} and $\tilde{\Sigma}$ are the same. However, these two covariance structures are generally different when we consider the regression problem (Kobayashi & Komaki, 2008; George & Xu, 2008). We extend the results for this general situation to matrix-variate case. We define

$$\Sigma_1 = \left\{ (\Sigma \otimes C)^{-1} + (\tilde{\Sigma} \otimes \tilde{C})^{-1} \right\}^{-1},$$

$$\Sigma_2 = \Sigma \otimes C,$$

and write the diagonalization of $\Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2}$ as

$$\Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} = U^\top \Lambda U,$$

where U is an orthogonal matrix and Λ is a diagonal matrix. Let

$$A^* = \Sigma_1^{1/2} U^\top (\Lambda^{-1} - I)^{1/2}.$$

From Theorem 3.2 of Kobayashi & Komaki (2008), we obtain the following Theorem.

Theorem 5. *If $\pi\{A^* \text{vec}(M)\}$ is superharmonic as a function of $\text{vec}(M)$, $\hat{p}_\pi(\tilde{Y}|Y)$ dominates $\hat{p}_I(\tilde{Y}|Y)$ under R_{KL}*

We can construct a prior that satisfies the condition of Theorem 5 by using the singular value shrinkage prior:

$$\pi(M) = \pi_{\text{SVS}} \left[\text{vec}^{-1} \left\{ (A^*)^{-1} \text{vec}(M) \right\} \right]. \quad (16)$$

The analytical form of Bayesian predictive densities based on prior (16) is not yet obtained. We conjecture that some extension of generalized Laguerre polynomial of matrix argument is necessary to solve this problem.

3 Application to the multivariate linear regression problems

In this section, we apply the results in the previous section to the multivariate linear regression problems.

In some areas, such as econometrics and chemometrics, linear regression problems often arise, wherein we have several dependent variables (Breiman & Friedman, 1997). Such problems are called multivariate linear regression. Namely,

$$Y \sim N_{n,q}(XB, \sigma^2 I, I), \quad (17)$$

where $X(n \times p)$ is a matrix of explanatory variables, $Y(n \times q)$ is a matrix of response variables and $B(p \times q)$ is a matrix of regression coefficients. We assume that σ^2 is known and $p \geq q$ in the following.

Let

$$Y_1 = (X^\top X)^{-1} X^\top Y,$$

and

$$Y_2 = Y - XY_1.$$

Then, we can reduce (17) to

$$\begin{aligned} Y_1 &\sim N_{p,q}(B, \sigma^2 (X^\top X)^{-1}, I), \\ Y_2 &\sim N_{n,q}(0, \sigma^2 I_{n-d}, I). \end{aligned} \quad (18)$$

Therefore, we can formulate multivariate linear regression to the form of prediction as

$$\begin{aligned} Y_1 &\sim N_{p,q}(B, \sigma^2 (X^\top X)^{-1}, I), \\ \tilde{Y}_1 &\sim N_{n,q}(B, \tilde{\sigma}^2 (\tilde{X}^\top \tilde{X})^{-1}, I). \end{aligned}$$

Thus we can reduce the multivariate linear regression problem to the matrix-variate Normal model.

Therefore, letting $C = \sigma^2 (X^\top X)^{-1}$, $\tilde{C} = \tilde{\sigma}^2 (\tilde{X}^\top \tilde{X})^{-1}$ and $\Sigma = \tilde{\Sigma} = I$, Bayesian predictive densities based on the prior (16) dominate those based on the Jeffreys prior. When constructing the predictive density of \tilde{Y} , we construct Bayesian predictive densities for \tilde{Y}_1 based on prior (16), and for \tilde{Y}_2 we use the density of (18) directly.

For the multivariate linear regression problem, it is useful to assume that the regression coefficient matrix B is low rank. This method is called reduced-rank regression (Reinsel & Velu, 1998). The rank of a matrix equals the number of non-zero singular values, so a low rank matrix has sparse singular values. Therefore, Bayesian predictive densities based on the singular value shrinkage prior (16) work particularly well in such situations. Also, they are minimax, whereas reduce-rank method is not minimax.

We can also apply this result to time series analysis. Reduced-rank methods for time series analysis are discussed in Velu et al. (1986). One typical example is the AR model for multivariate time series:

$$x_t = \sum_{k=1}^p A_k x_{t-k} + \varepsilon_t, \quad \varepsilon_t \stackrel{\text{i.i.d.}}{\sim} N_d(0, \Sigma).$$

Often we can assume that $A = (A_1 A_2 \cdots A_p)$ is reduced-rank. Therefore we obtain a better prediction of future variables by using the singular value shrinkage priors.

4 Numerical results

In this section, we show several experimental results on the Bayesian inference with singular value shrinkage priors. We used two methods to calculate the hypergeometric function of matrix argument in (11). First is an algorithm by Koev & Edelman (2006) that exploits the combinatorial properties of the Jack function. Second is the holonomic gradient method of Hashiguchi et al. (2013), wherein we numerically solve a PDE that ${}_1F_1$ satisfies. While the first method does not work well for matrix arguments with large singular values, the second method can in principle be used for any matrix arguments.

We compare the singular value shrinkage prior to the Jeffreys prior and the Stein prior. Here, the Jeffreys prior coincides with the uniform prior and the Stein prior is

$$\pi_S(M) = \|M\|_F^{-(nm-2)},$$

where $\|M\|_F$ denotes the Frobenius norm of M . We note that

$$\|M\|_F^2 = \sum_{i=1}^m \sigma_i(M)^2,$$

where $\sigma_i(M)$ is the i th singular value of M , holds.

First, we investigate the performance of Bayes estimators based on the singular value shrinkage priors.

We comment on the calculation method. From Stein (1974), Bayes estimator for the mean matrix parameter with prior $\pi(M)$ is

$$\left\{ \widehat{M}^\pi(X) \right\}_{ij} = X_{ij} + \frac{\partial}{\partial X_{ij}} \log m_\pi(X),$$

where $m_\pi(X)$ is the marginal distribution of X with prior $\pi(M)$. When π is the singular value shrinkage prior π_{SVS} , by (14),

$$\frac{\partial}{\partial X_{ij}} \log m_{\text{SVS}}(X) = -X_{ij} + \frac{\partial}{\partial X_{ij}} \left\{ \log {}_1F_1 \left(\frac{m+1}{2}; \frac{n}{2}; \frac{1}{2} X^\top X \right) \right\}.$$

In the numerical experiment, we approximated the partial differentiation of $\log {}_1F_1$ by finite difference:

$$\begin{aligned} & \frac{\partial}{\partial X_{ij}} \left\{ \log {}_1F_1 \left(\frac{m+1}{2}; \frac{n}{2}; \frac{1}{2} X^\top X \right) \right\} \\ & \approx (2\varepsilon)^{-1} \left\{ \log {}_1F_1 \left(\frac{m+1}{2}; \frac{n}{2}; \frac{1}{2} (X + \varepsilon E^{(i,j)})^\top (X + \varepsilon E^{(i,j)}) \right) \right. \\ & \quad \left. - \log {}_1F_1 \left(\frac{m+1}{2}; \frac{n}{2}; \frac{1}{2} (X - \varepsilon E^{(i,j)})^\top (X - \varepsilon E^{(i,j)}) \right) \right\}, \end{aligned}$$

where $E^{(i,j)}$ is a $n \times m$ matrix whose (i, j) th element is 1 and other elements are 0. We put $\varepsilon = 10^{-6}$.

In the following, we compare the risk functions of the Bayes estimator with the Jeffreys prior (it coincides with the maximum likelihood estimator), the Bayes estimator with the Stein prior, and the Bayes estimator with the singular value shrinkage prior. We sampled X 10^4 times and approximated the risk by the sample mean of loss (square of the Frobenius norm of the difference between estimate and the true parameter).

Figure 1 shows the risk functions for $m = 2$, $n = 4$ and $\sigma_1 = 20$. The singular value shrinkage prior performs better than the Jeffreys prior, and the amount of risk reduction increases as σ_2 decreases. On the other hand, the Stein prior does not perform well because the Frobenius norm of A is not small, even when $\sigma_2 = 0$.

Figure 2 shows the risk functions for $m = 2$, $n = 4$ and $\sigma_2 = 0$. Though the Stein prior performs best when σ_1 is small, the risk of the Stein prior becomes almost the same as the maximum likelihood estimator as σ_1 increases. On the other hand, the singular value shrinkage prior performs stably better than the maximum likelihood estimator regardless of the value of σ_1 . This is because the singular value shrinkage prior shrinks σ_1 and σ_2 for each, while the Stein prior shrinks $\sigma_1^2 + \sigma_2^2$, the Frobenius norm.

Figure 3 shows the risk functions for $m = 3$, $n = 5$ and $\sigma_1 = 5$, $\sigma_3 = 0$ and Figure 4 shows the risk functions for $m = 3$, $n = 5$ and $\sigma_2 = \sigma_3 = 0$. The performance of the singular value shrinkage prior is qualitatively the same as when $m = 2$, $n = 4$. Figure 4 indicates that this prior works stably well for low rank matrices.

Next, we compare the risk of Bayesian predictive densities based on the Jeffreys prior, the Stein prior and singular value shrinkage prior. We sampled X 10^4 times and approximated the risk by the sample mean of the Kullback–Leibler loss.

Figure 5-8 shows the risk functions for the same settings in Figure 1-4. The performance of singular value shrinkage prior is qualitatively the same as in the estimation.

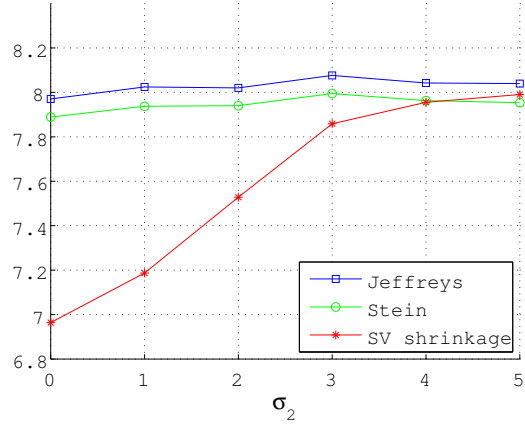


Figure 1: Risk function of Bayes estimators. $m = 2$, $n = 4$ and $\sigma_1 = 20$.

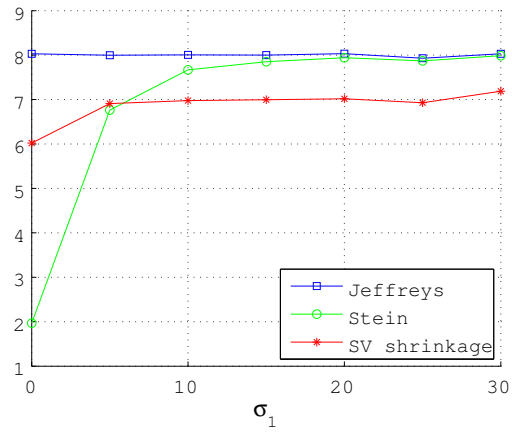


Figure 2: Risk function of Bayes estimators. $m = 2$, $n = 4$ and $\sigma_2 = 0$.

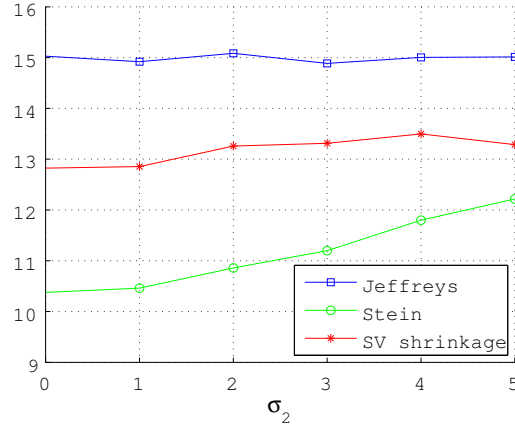


Figure 3: Risk function of Bayes estimators. $m = 3$, $n = 5$ and $\sigma_1 = 5$, $\sigma_3 = 0$.

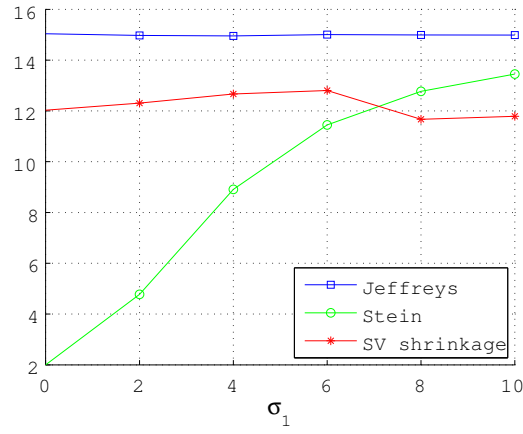


Figure 4: Risk function of Bayes estimators. $m = 3$, $n = 5$ and $\sigma_2 = \sigma_3 = 0$.

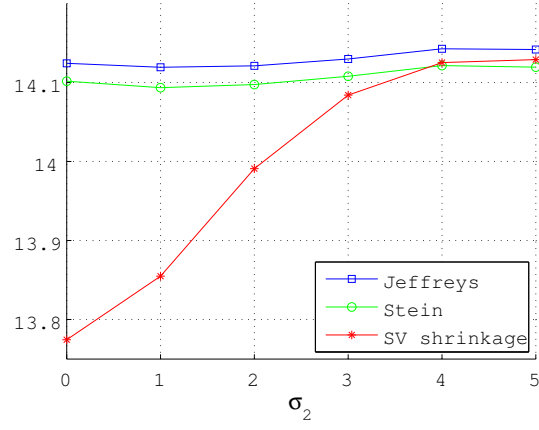


Figure 5: Risk function of predictive densities. $m = 2$, $n = 4$ and $\sigma_1 = 20$.

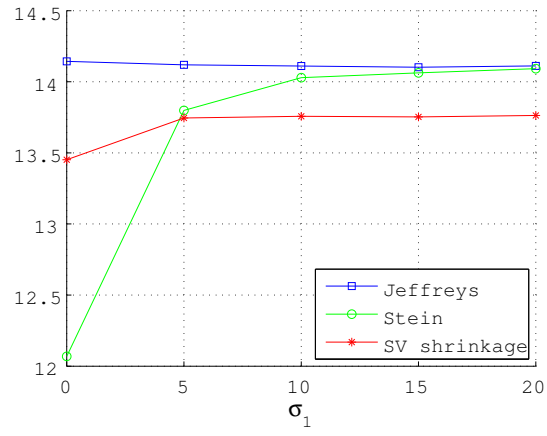


Figure 6: Risk function of predictive densities. $m = 2$, $n = 4$ and $\sigma_2 = 0$.

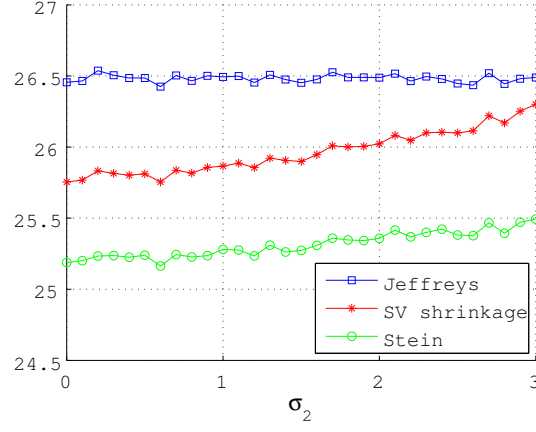


Figure 7: Risk function of predictive densities. $m = 3$, $n = 5$ and $\sigma_1 = 5, \sigma_3 = 0$.

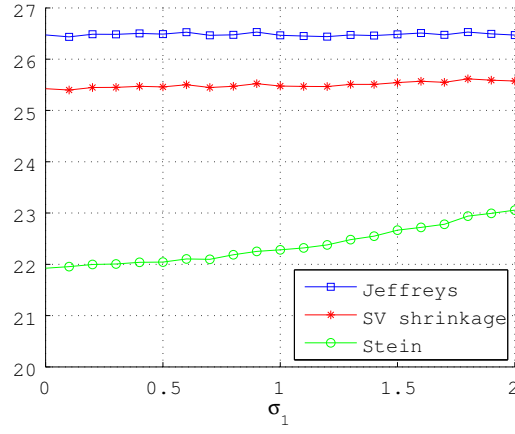


Figure 8: Risk function of predictive densities. $m = 3$, $n = 5$ and $\sigma_2 = \sigma_3 = 0$.

5 Concluding remarks

In this paper, we developed singular value shrinkage priors for the mean matrix parameters of the matrix-variate Normal model. We proved the superharmonicity of the introduced prior. From the previous studies about the relationship between the superharmonicity of a prior and the minimaxity of the Bayes estimator and the Bayesian predictive density, Bayes estimators and Bayesian predictive densities based on the introduced prior are minimax and also dominate those based on the Jeffreys prior. We provided the analytical form of Bayesian predictive densities based on the singular value shrinkage prior, using the classical results in multivariate analysis. Next, we applied the introduced prior to multivariate regression problems. Here we considered priors depending on the future samples. In multivariate regression problems, the matrix of regression coefficients is often assumed to be low rank. Singular value shrinkage prior works stably well in such situations. Finally, numerical results showed the effectiveness of singular value shrinkage priors.

We present several problems that require further study.

We obtained the analytical form of Bayesian predictive densities based on singular value shrinkage priors for the case where $\tilde{\Sigma} \otimes \tilde{C}$ is proportional to $\Sigma \otimes C$. The calculation for the general covariance case is a future problem. This is indispensable to obtain Bayesian predictive densities in the multivariate linear regression problems. We conjecture that some extension of generalized Laguerre polynomial of matrix argument is necessary to solve this problem.

We extended the Bayesian shrinkage prediction method for vector-variate Normal distributions of Komaki (2001) and George et al. (2006) to matrix-variate case. Then, we can consider further extensions to tensor-variate case. Recently, tensor data has gained popularity (Kolda & Bader, 2009). However, there are several technical problems with this extension. First, we have difficulties in defining singular values for tensors. Although Lathauwer et al. (2000) provides one definition of singular values for tensors, it does not seem to be a definitive answer. Second, since it is uncommon that all entries of tensor data can be observed, we should consider prediction with observing only limited entries. Finally, the method of regression with tensor is now emerging in areas such as brain signal processing. We expect that the singular value shrinkage prior works well in this problem as well as in multivariate linear regression.

Appendix The singular value coordinate system

Here, we introduce a coordinate system on the space of matrices that is useful for treating singular values. We put the singular value decomposition

of a matrix X as

$$\begin{aligned} X &= U\Sigma V^\top, \\ U &\in O(n), \quad V \in O(m), \quad \Sigma = (\text{diag}(\sigma_1, \dots, \sigma_m) \quad O_{m, n-m})^\top, \\ \sigma_1 &\geq \sigma_2 \geq \dots \geq \sigma_m \geq 0, \end{aligned} \tag{19}$$

where $O_{m, n-m}$ is the zero matrix of size $m \times (n - m)$ and $\sigma_1, \dots, \sigma_m$ are the singular values of X .

Since any orthogonal matrix can be represented as the exponential of an anti-symmetric matrix, there exist matrices A ($m \times m$) and B ($n \times n$) such that

$$\begin{aligned} U &= \exp A, \quad A^\top = -A, \\ V &= \exp B, \quad B^\top = -B. \end{aligned}$$

We put the entries of A and B as

$$\begin{aligned} A &= \begin{pmatrix} 0 & \alpha_{1,2} & \alpha_{1,3} & \cdots & \alpha_{1,m-1} & \alpha_{1,m} \\ -\alpha_{1,2} & 0 & \alpha_{2,3} & \cdots & \alpha_{2,m-1} & \alpha_{2,m} \\ -\alpha_{1,3} & -\alpha_{2,3} & 0 & \cdots & \alpha_{3,m-1} & \alpha_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\alpha_{1,m-1} & -\alpha_{2,m-1} & -\alpha_{3,m-1} & \cdots & 0 & \alpha_{m-1,m} \\ -\alpha_{1,m} & -\alpha_{2,m} & -\alpha_{3,m} & \cdots & -\alpha_{m-1,m} & 0 \end{pmatrix}, \\ B &= \begin{pmatrix} 0 & \beta_{1,2} & \beta_{1,3} & \cdots & \beta_{1,n-1} & \beta_{1,n} \\ -\beta_{1,2} & 0 & \beta_{2,3} & \cdots & \beta_{2,n-1} & \beta_{2,n} \\ -\beta_{1,3} & -\beta_{2,3} & 0 & \cdots & \beta_{3,n-1} & \beta_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\beta_{1,n-1} & -\beta_{2,n-1} & -\beta_{3,n-1} & \cdots & 0 & \beta_{n-1,n} \\ -\beta_{1,n} & -\beta_{2,n} & -\beta_{3,n} & \cdots & -\beta_{n-1,n} & 0 \end{pmatrix}. \end{aligned}$$

Then we can use (α, σ, β) as a local coordinate, where $\alpha = (\alpha_{1,2}, \dots, \alpha_{m-1,m})$ represents the entries of A , $\sigma = (\sigma_1, \dots, \sigma_m)$ represents the singular values of X , and $\beta = (\beta_{1,2}, \dots, \beta_{m,n})$ represents the entries of the first to m th row of B .

It is proved as follows. Let \tilde{V} be the matrix composed of the first to m th row of V and $\tilde{\Sigma}$ be the matrix composed of the first to m th column of Σ . Thus,

$$\tilde{V} \in \mathbb{R}^{m \times n}, \quad \tilde{V}\tilde{V}^\top = I,$$

and

$$\tilde{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_m).$$

Then, (19) can be rewritten as

$$X = U\tilde{\Sigma}\tilde{V}^\top. \tag{20}$$

Now, in the neighborhood of $B = 0$, the first to m th row of V depend only on the first to m th row of B , since

$$V = \exp B = I + B + O(B^2).$$

Therefore, (α, σ, β) can be used as a local coordinate system.

Next, we calculate the metric tensor and the Laplacian form in this coordinate system.

The metric tensor does not depend on A and B from invariance. Therefore we calculate the metric tensor on the neighborhood of $A = 0, B = 0$. From (20), we have

$$\begin{aligned} dX &= (\exp A \cdot dA) \cdot \Sigma (\exp B)^\top + \exp A \cdot d\Sigma \cdot (\exp B)^\top \\ &\quad + \exp A \cdot \Sigma \cdot (\exp B)^\top \cdot dB^\top. \end{aligned}$$

Putting $A = 0, B = 0$, we get

$$dX = dA \cdot \Sigma + d\Sigma + \Sigma \cdot dB.$$

Then, the components of dX are obtained as follows:

$$\begin{aligned} (dX)_{ii} &= d\sigma_i, \\ (dX)_{ij} &= -\sigma_j d\alpha_{ji} + \sigma_i d\beta_{ji} \quad \text{if } i > j, \\ (dX)_{ij} &= \sigma_j d\alpha_{ij} - \sigma_i d\beta_{ij} \quad \text{if } i < j, \quad j \leq m, \\ (dX)_{ij} &= -\sigma_i d\beta_{ij} \quad \text{if } i < j, \quad m < j. \end{aligned}$$

Therefore,

$$\begin{aligned} ds^2 &= \sum_{i=1}^m \sum_{j=1}^n (dX)_{ij}^2 \\ &= \sum_{i=1}^m d\sigma_i^2 + \sum_{i=1}^m \sum_{j=i+1}^m \{(\sigma_i^2 + \sigma_j^2)(d\alpha_{ij}^2 + d\beta_{ij}^2) - 4\sigma_i \sigma_j d\alpha_{ij} d\beta_{ij}\} \\ &\quad + \sum_{i=1}^m \sum_{j=m+1}^n \sigma_i^2 d\beta_{ij}^2 \end{aligned}$$

and it follows that

$$\begin{aligned} (g_{\alpha\alpha})_{(i,j)(i',j')} &= (\sigma_i^2 + \sigma_j^2) \delta_{(i,j)(i',j')}, \\ (g_{\sigma\sigma})_{ij} &= \delta_{ij}, \\ (g_{\beta\beta})_{(i,j)(i',j')} &= (\sigma_i^2 + \sigma_j^2) \delta_{(i,j)(i',j')} \quad \text{if } j \leq m, \\ (g_{\beta\beta})_{(i,j)(i',j')} &= \sigma_i^2 \delta_{(i,j)(i',j')} \quad \text{if } j > m, \\ (g_{\alpha\sigma})_{(i,j)(i',j')} &= (g_{\sigma\beta})_{(i,j)(i',j')} = 0, \\ (g_{\alpha\beta})_{(i,j)(i',j')} &= -2\sigma_i \sigma_j \delta_{(i,j)(i',j')} \quad \text{if } j \leq m, \\ (g_{\alpha\beta})_{(i,j)(i',j')} &= 0 \quad \text{if } j > m. \end{aligned}$$

In the matrix form, the metric tensor can be written as

$$g = \begin{pmatrix} g_{\alpha\alpha} & g_{\alpha\sigma} & g_{\alpha\beta} \\ g_{\sigma\alpha} & g_{\sigma\sigma} & g_{\sigma\beta} \\ g_{\beta\alpha} & g_{\beta\sigma} & g_{\beta\beta} \end{pmatrix} = \begin{pmatrix} g_{\alpha\alpha} & O & g_{\alpha\beta} \\ O & I_m & O \\ g_{\beta\alpha} & O & g_{\beta\beta} \end{pmatrix}, \quad (21)$$

where I_m is the m -dimensional identity matrix.

Then we can calculate the Laplacian form in this coordinate. The Laplacian form on Riemannian manifolds can generally be written as

$$\Delta f = \frac{1}{\sqrt{\det(g_{ij})}} \frac{\partial}{\partial u^i} \left(\sqrt{\det(g_{ij})} g_{ij} \frac{\partial f}{\partial u^j} \right) \quad (22)$$

where $\det(g_{ij})$ is the determinant of the metric tensor.

By applying the formula for the determinant of a block matrix

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det A \cdot \det(D - CA^{-1}B)$$

to (21), we obtain

$$\det(g_{ij}) = \det \begin{pmatrix} g_{\alpha\alpha} & g_{\alpha\sigma} \\ g_{\sigma\alpha} & g_{\sigma\sigma} \end{pmatrix} \det \left\{ g_{\beta\beta} - (g_{\beta\alpha} \ g_{\beta\sigma}) \begin{pmatrix} g_{\alpha\alpha} & g_{\alpha\sigma} \\ g_{\sigma\alpha} & g_{\sigma\sigma} \end{pmatrix}^{-1} \begin{pmatrix} g_{\alpha\beta} \\ g_{\sigma\beta} \end{pmatrix} \right\}.$$

Then, from

$$\det \begin{pmatrix} g_{\alpha\alpha} & g_{\alpha\sigma} \\ g_{\sigma\alpha} & g_{\sigma\sigma} \end{pmatrix} = \det \begin{pmatrix} g_{\alpha\alpha} & O \\ O & I_m \end{pmatrix} = \det(g_{\alpha\alpha}) = \prod_{i < j}^m (\sigma_i^2 + \sigma_j^2),$$

$$(g_{\beta\alpha} \ g_{\beta\sigma}) \begin{pmatrix} g_{\alpha\alpha} & g_{\alpha\sigma} \\ g_{\sigma\alpha} & g_{\sigma\sigma} \end{pmatrix}^{-1} \begin{pmatrix} g_{\alpha\beta} \\ g_{\sigma\beta} \end{pmatrix} = \prod_{i < j}^m \left(\sigma_i^2 + \sigma_j^2 - \frac{4\sigma_i^2 \sigma_j^2}{\sigma_i^2 + \sigma_j^2} \right) \prod_{i=1}^m \sigma_i^{2(n-m)}$$

we have

$$\begin{aligned} |g| &= \prod_{i < j}^m (\sigma_i^2 + \sigma_j^2) \cdot \prod_{i < j}^m \left(\sigma_i^2 + \sigma_j^2 - \frac{4\sigma_i^2 \sigma_j^2}{\sigma_i^2 + \sigma_j^2} \right) \prod_{i=1}^m \sigma_i^{2(n-m)} \\ &= \prod_{i < j}^m (\sigma_i^2 - \sigma_j^2)^2 \cdot \prod_{i=1}^m \sigma_i^{2(n-m)}. \end{aligned}$$

Therefore, the Laplacian form (22) can be written as

$$\Delta f = 2 \sum_{i < j} \frac{\sigma_i \frac{\partial f}{\partial \sigma_i} - \sigma_j \frac{\partial f}{\partial \sigma_j}}{\sigma_i^2 - \sigma_j^2} + (n-m) \sum_{i=1}^m \frac{1}{\sigma_i} \frac{\partial f}{\partial \sigma_i} + \sum_{i=1}^m \frac{\partial^2 f}{\partial \sigma_i^2}. \quad (23)$$

From (23) we can determine whether a function $f(\alpha, \sigma, \beta)$ is superharmonic or not. For example, the Laplacian of π_{SVS} is

$$\begin{aligned}\Delta\pi_{\text{SVS}} &= 2 \sum_{i < j} \frac{\sigma_i \frac{-(n-m-1)f}{\sigma_i} - \sigma_j \frac{-(n-m-1)f}{\sigma_j}}{\sigma_i^2 - \sigma_j^2} \\ &\quad - (n-m) \sum_{i=1}^m \frac{1}{\sigma_i} \frac{(n-m-1)f}{\sigma_i} + \sum_{i=1}^m \frac{(n-m)(n-m-1)f}{\sigma_i^2} \\ &= 0.\end{aligned}$$

This method of determining the superharmonicity of a prior is useful for various forms of priors that can be represented with singular values. We can construct priors with different strengths of shrinkage for each singular value, and determine whether it is superharmonic.

We can infer the explicit form of the prior π_{SVS} naturally as follows. First, we note the relationship between the James-Stein estimator and the Stein prior. The James-Stein estimator is

$$\hat{\mu}_{\text{JS}} = \left(1 - \frac{d-2}{\|x\|^2}\right) x,$$

and the Stein prior is

$$\pi_{\text{S}}(\mu) = \|\mu\|^{-(d-2)}.$$

Here, we notice that $(d-2)$ appears in common.

Now, the shrinkage estimator of Efron & Morris (1972) is

$$\hat{M}_{\text{EM}} = X \{I_m - (n-m-1)S^{-1}\} \quad (i = 1, \dots, m),$$

which can be represented in singular value coordinate as

$$\hat{\sigma}_i = \left(1 - \frac{n-m-1}{\sigma_i^2}\right) \sigma_i.$$

Then, from the analogy above, we can infer the prior form of

$$\pi_{\text{SVS}}(\alpha, \sigma, \beta) = \prod_{i=1}^m \sigma_i^{-(n-m-1)},$$

where the prior for α, β is uniform in the same way as the Stein prior depends only on the norm and not on the angle.

References

BREIMAN, L. & FRIEDMAN, J. H. (1997). Predicting multivariate responses in multiple linear regression. *J. R. Statist. Soc. B* **59**, 3–54.

- DAWID, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68**, 265–74.
- EFRON, B. & MORRIS, C. (1972). Empirical Bayes on vector observations: an extension of Stein’s method. *Biometrika* **59**, 335–47.
- GEORGE, E. I., LIANG, F. & XU, X. (2006). Improved minimax predictive densities under Kullback–Leibler loss. *Ann. Statist.* **34**, 78–91.
- GEORGE, E. I. & XU, X. (2008). Predictive density estimation for multiple regression. *Economet. Theor.* **24**, 528–44.
- GUPTA, A. K. & NAGAR, D. K. (2000). *Matrix variate distributions*. New York: Chapman & Hall.
- HASHIGUCHI, H., NUMATA, Y., TAKAYAMA, N. & TAKEMURA, A. (2013). The holonomic gradient method for the distribution function of the largest root of a Wishart matrix. *J. Multivariate. Anal.* **117**, 296–312.
- HELMS, L. L. (2009). *Potential Theory*. New York: Springer-Verlag.
- KOBAYASHI, K. & KOMAKI, F. (2008). Bayesian shrinkage prediction for the regression problem. *J. Multivariate. Anal.* **99**, 1888–1905.
- KOEV, P. & EDELMAN, A. (2006). The efficient evaluation of the hypergeometric function of a matrix argument. *Math. Comput.* **75**, 833–46.
- KOLDA, T. G. & BADER, B. W. (2009). Tensor decompositions and applications. *SIAM. Rev.* **51**, 455–500.
- KOMAKI, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika* **88**, 859–64.
- LATHAUWER, L., MOOR, B. & VANDEWALLE, J. (2000). A multilinear singular value decomposition. *SIAM. J. Matrix. Anal. A.* **21**, 1253–78.
- REINSEL, G. C. & VELU, R. P. (1998). *Multivariate Reduced-Rank Regression*. New York: Springer-Verlag.
- STEIN, C. (1974). Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symposium on Asymptotic Statistics* **2**, 345–81.
- VELU, R. P., REINSEL, G. C. & WICHERN, D. W. (1986). Reduced rank models for multiple time series. *Biometrika* **73**, 105–18.